BIG DATA ANALYTICS
SPARK SQL AND DATAFRAMES
HOMEWORK – 6

Q1. What is Spark SQL?

Spark SQL is a spark module for structured data processing. Spark SQL integrates relational processing with Spark's functional programming. It provides support for various data sources and makes it possible to weave SQL queries with code transformations thus resulting in a very powerful tool.

Q2. Is there a module to implement SQL in Spark?

How does it work?

Spark SQL is a new module in Spark which integrates relational processing with Spark's functional programming API. It supports querying data either via SQL or via the Hive Query Language.

Spark SQL blurs the line between RDD and relational table. It offers much tighter integration between relational and procedural processing, through declarative DataFrame APIs which integrates with Spark code. It also provides higher optimization. DataFrame API and Datasets API are the ways to interact with Spark SQL.

Q3. What is a Parquet File?

Parquet is a columnar format and Spark SQL provides support for both reading and writing parquet files that automatically preserves the schema of the original data. When reading Parquet files, all columns are automatically converted to be nullable for compatibility reasons.

Q4. List the functions of Spark SQL.

Spark SQL is capable of:

- Loading data from a variety of structured sources
- Querying data using SQL statements, both inside a Spark program and from external tools that connect to Spark SQL through standard database connectors (JDBC/ODBC). For instance, using business intelligence tools like Tableau
- Providing rich integration between SQL and regular Python/Java/Scala code, including the ability to join RDDs and SQL tables, expose custom functions in SQL, and more

Q5. How is Spark SQL different from HQL and SQL?

Spark SQL is a special component on the spark Core engine that support SQL and Hive Query Language without changing any syntax. It's possible to join SQL table and HQL table.

Q6. Why is Spark SQL used?

Spark SQL originated as Apache Hive to run on top of Spark and is now integrated with the Spark stack. Apache Hive had certain limitations as mentioned below. Spark SQL was built to overcome these drawbacks and replace Apache Hive.

Q7. Is Spark SQL faster than Hive?

Spark SQL is faster than Hive when it comes to processing speed.

Limitations With Hive:

- Hive launches MapReduce jobs internally for executing the ad-hoc queries. MapReduce lags in the performance when it comes to the analysis of medium-sized datasets (10 to 200 GB).

- Hive has no resume capability. This means that if the processing dies in the middle of a workflow, you cannot resume from where it got stuck.
- Hive cannot drop encrypted databases in cascade when the trash is enabled and leads to an execution error. To overcome this, users have to use the Purge option to skip trash instead of drop.

These drawbacks gave way to the birth of Spark SQL.